科学基金资助机构视角下的科学数据管理研究

赵秋红[*]　李元睿　邓修权　张 楚　张保丰

北京航空航天大学 经济管理学院 北京 100191

摘要　随着科学数据的获取、存储、分析和处理等技术的发展，科学研究与科技创新逐步走向以科学数据为基础性科技资源的大数据时代，数据驱动研究范式渗入到各学科领域的实际工作中，科学数据在科研创新中的价值日益突出，科学基金资助机构的科学数据管理职责也更加重要。为此，文章在分析科学数据管理需求产生的驱动因素的基础之上，梳理了发达国家科学数据管理的实践经验，指出科学数据管理活动应联结科学数据全生命周期各阶段，以扩展科学数据的全生命周期及其产生的价值、推动其健康和可持续发展。文章提出了针对科学数据全生命周期的管理策略，包括：科学数据管理计划的制定和实施、科学数据汇交管理、科学数据开放共享、科学数据的可持续维护等，并围绕这些管理策略提出了具体的实施建议。

关键词　科学数据，科学基金资助机构，数据汇交，数据共享

DOI　10.16418/j.issn.1000-3045.20211116001

科学数据是指将研究对象抽象化和概念化后所形成的、用于科学研究活动的相关事实记录[1]。科学数据的形式包括统计数据、实验结果、观测结果、访谈记录、图像和声音等，是证实科研发现或支撑学术观点的证据，也是进行理论推理的基础[2]。随着大数据时代的到来，海量科学数据带来了丰富的基础性科技资源，科学研究水平逐渐开始依赖于对科学数据的积累，以及将科学数据转化为知识和科研产出的能力。科学数据管理是指协调并规范对科学数据的采集、生产、存储、使用、共享等活动。对科学基金资助机构而言，科学数据管理就是对被资助者设定并监督其科学数据采集、生产、汇交的标准和流程，加强科学数据存储与共享的软件系统与硬件设施建设，推动被资助项目所产生的科学数据的开放共享，发挥科学数据所蕴含的价值[3]。

当前，在国家科学数据管理政策的指导下，一

*通信作者

资助项目：国家自然科学基金应急管理项目（71843011）
修改稿收到日期：2021年11月24日

些部门特别是国家级科学数据中心制定了具体的科学数据管理方案，有效开展了科学数据管理的实践 [4]。但科学基金资助机构尚未形成成熟的科学数据管理方案。科学基金资助机构是资助科学研究的主要渠道之一，对科学基金资助机构的科学数据管理具有重要意义：一方面，科学基金资助机构承担着所资助项目的管理责任，如何把所资助项目产生的数据收集好、管理好、利用好，是一项重要的任务；另一方面，科学基金资助机构掌握大量的科学数据资源和相关信息，与资助方、项目承担者及其所在工作单位存在紧密和长期的合作关系，具有科学数据管理的先天条件和优势。因此，提高科学基金资助机构的科学数据管理水平，进一步促进科学数据开放共享，是提升我国科研水平和创新能力的重要途径，具有重要的战略意义。

1 科学数据管理需求产生的驱动因素

1.1 科学研究范式转变

在科学史中，无论“经验范式”“理论范式”或“计算范式”，用数据研究科学规律始终扮演着重要的角色。随着信息技术革命的发展，科学数据越来越容易被生产（收集）、存储、处理、分析和传播，科学数据总量呈几何式增长，这使得任何单一的传统研究范式都无法有效应对密集型数据的挖掘和整合 [5]。因此，科学研究范式开始转向“第四范式”，即“数据驱动范式” [6]。在此背景下，学科交叉融合与科学数据爆炸式增长相互促进，科学数据管理越发成为整合数据资源的必要手段。

1.2 大数据时代推动

近年来，高度连接的世界和迅速发展的电子信息相关的软、硬件设备使得数据产生的范围、方式、途径发生了革命性变化。数据在类型格式、组成结构、存在形态等方面也趋向复杂化[7]。在云计算、大数据分析工具、并行数据库等技术工具的支撑下，从海量数据中挖掘出新的知识变为可能，科学数据越发成为科学研究的“金矿”；围绕科学数据的存储、分析、传播和应用等要素的科学数据管理正越来越影响着一个国家的科技水平。

1.3 开放获取运动兴起

开放获取（Open Access）是致力于推动科研成果共享，借助互联网自由传播的特性来促进科研交流，推动便捷出版，提高科研效率的行动[8]。在数据资源领域，科学数据的开放共享能够减少重复劳动，缩短科研周期。然而，在复杂的科研场景下，数据的展现形式和获取途径难以满足知识共同体的需求，需要科学的激励机制和质量控制体系来保证科学数据的有效流动，从而形成博弈策略的稳态平衡。因此，实施科学数据管理也是开放获取运动的必然要求。

除了上述因素以外，不断扩张的科学数据边界、数据结构多样性、数据权益及数据隐私保护等因素也是驱动科学数据管理不断发展的重要因素。因此，各国政府对科学数据资源提高重视，不断加强政策引导以推动数据开放共享。

2 部分发达国家科学资助机构的科学数据管理实践

2.1 美国主要科学资助机构的科学数据管理实践

美国国家科学基金会（NSF）要求所资助的科研项目在项目申请阶段应提交“数据管理计划”（Data Management Plan，DMP），以加强对所资助科研项目产出的科学数据的管理。在 DMP 中，项目申请人需要对项目实施中产生的所有科学数据及其元数据的格式、内容标准、访问权限、共享计划等内容进行阐述。该计划是项目审核的先决条件和重要评判依据。美国国立卫生研究院（NIH）同样制定了科学数据管理相关政策，并要求项目产生的科研数据要符合 FAIR 原则[9]，即：可检索（findable）、可访问（accessible）、可交互使用（interoperable）和可重复

使用（reusable）。

2.2 英国主要科学资助机构的科学数据管理实践

英国研究理事会（RCUK）等科学资助机构是英国科学数据管理政策的主要制定者。RCUK发布了多项科学数据管理政策，提出了包括数据成长、长期存储、共享和开放等方面的数据管理政策的基本原则，指出科学数据管理需要遵循的5项原则：① 明确研究人员、研究机构和资助者的责任和义务；② 在收集和筛选科学数据时，应保证数据质量；③ 数据共享时应提高科学数据的查询效率，提供访问的权限；④ 科学制定科学数据管理政策办法，提高公共科研基金的使用效率和使用效益；⑤ 对具有长期价值的科学数据进行妥善保存。

2.3 澳大利亚主要资助机构的科学数据管理实践

澳大利亚国家数据服务局（ANDS）为科研工作者提供数据管理服务，以致力于提高科学数据的价值。ANDS对科学数据管理中需要考虑的关键步骤进行了研究，并明确了在这些步骤下的责任划分。澳大利亚研究理事会（ARC）、澳大利亚国家卫生和医学研究理事会（NHMRC）等科学资助机构认同ANDS所拟定的科学数据管理计划，要求所资助的项目遵循这些规定，并鼓励研究团队将项目产出的科学数据及出版物存储在指定的数据库中以便于开放共享。

3 基于全生命周期的科学基金资助机构科学数据管理思路

基于上述分析，部分发达国家科学数据管理实践注重从宏观角度把握科学数据生命周期内的各项管理环节，尤其围绕DMP展开对科学数据从产生到再利用的各阶段的把关控制[10]。以DMP为抓手的管理思路值得我国科学数据管理实践借鉴。然而，现行科学数据管理活动往往限制于数据生命周期的各个离散的发展阶段中，不利于将各阶段有机联结[11]。

根据科学数据生产前、中、后3个时期，可将科学数据的生命周期划分为：数据的计划、数据的生成/收集、数据的处理、数据的存储、数据的共享、数据的再利用6个阶段。由于科学数据的产生和应用具有连续性特征，需要执行科学数据管理的环节不能完全与数据生命周期的各阶段一一对应。因此，本文提出全生命周期视角下的科学数据管理的总体思路（图1）。

全生命周期视角下的科学数据管理强调管理环节对数据生命周期各阶段的延伸影响和长期支持。在职责划分上，依托单位指导并管理科研团队，共同承担对科学数据产生前和产生中的任务，即制定并按照DMP产出和汇交符合要求的科学数据。第三方共享平台负责数据汇交完成后的数据存储、共享和再利用等阶段的工作。科学基金资助机构的职责贯穿数据的全生命周期：在科学数据的计划阶段，应推动DMP的制定、实施和考核；在数据产生的初期，应着手启动科学数据汇交管理，从软、硬件设施为科学数据汇交存储提供先决条件，并从利于数据共享的角度设计汇交流程和汇交模式；在数据产生后，应启动科学数据开放共享和科学数据的可持续维护，建立高效的共享机制，不断发掘数据的价值，直至科学数据过于陈旧，不再被人使用，即科学数据生命周期终结。

4 全生命周期视角下的科学基金资助机构科学数据管理方案

4.1 DMP的制定和实施

DMP的制定和实施应该在数据产生前和初期产生阶段进行，对应的是科研团队准备和提交项目申请书阶段。科学基金资助机构应要求科研团队提交详尽的DMP，并严格按照DMP对科学数据生命周期各阶段进行评估。项目团队需要通过DMP描述在项目研究过程中将要收集或产生的数据，并且明确在项目研究过程中如何管理和储存这些科学数据，以及在项

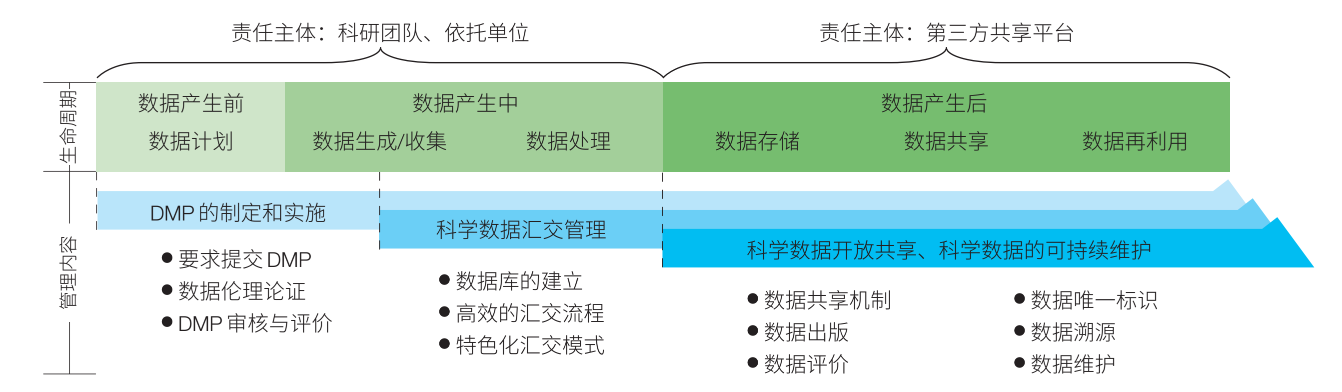


图 1　全生命周期视角下的科学数据管理环节

Figure 1　Phases of scientific data management during data life circle

目结题后如何共享。作为贯穿科学数据全生命周期的纲领，DMP 为数据伦理论证、追溯科学基金数据责任人提供了透明化路径和依据[13]。科学基金资助机构应建立先汇交项目科学数据、再验收项目的机制，将 DMP 执行的情况作为项目结题评审的重要考核指标，并把基于数据全生命周期的 DMP 执行情况作为申请新项目资质的条件。

4.2 科学数据汇交管理

科学数据的产生具有阶段性和持续性特点。因此，数据的汇交应发生在一定时间段内，即数据生成/收集到数据处理阶段。科学基金资助机构可设置灵活机制，允许科研团队随时将成熟的科学数据进行汇交，以提高数据的时效性，使其尽快被共享，提升其价值。应建立基于元数据的科学数据汇交模式（图 2）。元数据即“数据的数据”，是对科学数据进行规范化的描述数据[14]。科学数据产生后，以依托单位为数据汇交单元，将下属各项目团队的科学数据和元数据汇交至科学数据汇交共享平台；平台进行初步验收和分拣，将元数据进行存储，并将科学数据存放至相应学科的科学数据中心。各学科科学数据中心在科学基金资助机构的指导下对科学数据进行汇交质量评价。评价结果反馈后，对于不合格的数据，科学数据汇交共享平台应要求依托单位和项目团队进行整改和重新提交。

4.3 科学数据开放共享

科学数据进行汇交后，基于元数据的存储模式为科学数据的共享和再利用提供了便捷和开放的途径。科学数据的共享交由第三方进行，应平衡利益相关者的利益诉求，引导各方积极推动科学数据的开放共享。科学数据共享的利益相关者包括：政府、科学基金资助机构、科学数据中心、依托单位、数据生产者、数据使用者、同行评审专家、受试者和出版者[15]。科学数据的共享需要数据全生命周期利益相关者共同参与，可构建科学数据共享机制体系（图 3）；各利益相关者应该通过制度和利益协调来实现科学数据共享的目标。政府作为资金提供者和管理政策的顶层设计者，应该建立全方位的政策引导体系，规范监督和引导各利益相关者的行为；科学基金资助机构作为科学数据共享组织系统的中枢，应联结各利益相关者团结协作，与科学数据共享平台、数据出版商建立合作联盟的管理模式。

4.4 科学数据的可持续维护

科学数据的可持续维护贯穿数据的存储、共享和再利用阶段。对科学数据的可持续维护是数据全生命周期管理的重要组成，是实现科学数据不断发挥价值的重要保障。应建立面向用户的数据获取技术体系，构建高效合理的存储层次结构，对热数据进行缓存或预取，将冷数据迁移至低速存储设备，从而优化系统

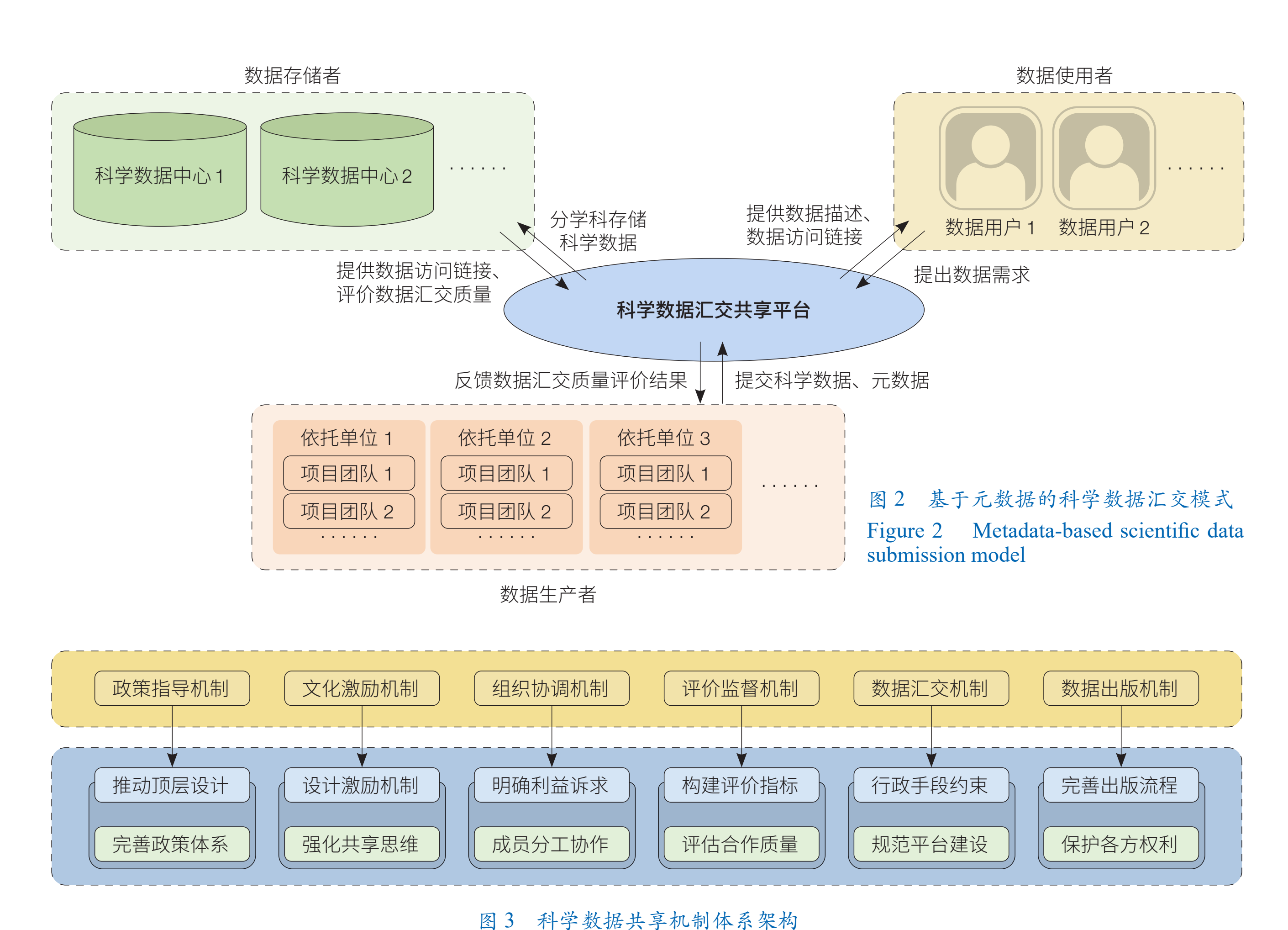


图 2　基于元数据的科学数据汇交模式

Figure 2　Metadata-based scientific data submission model

图 3　科学数据共享机制体系架构

Figure 3　Structure of scientific data sharing mechanism

性能分配，提高用户的数据获取便利程度。在数据溯源方面，应对提交的数据建立唯一标识，确保科学数据能够按照统一的标准进行整合，从而保证科学数据能够依据标识进行溯源，进一步确保科学数据可以被应用、比对。同时，建立基于身份证号码或开放研究者与贡献者身份（ORCID）的身份标识体系，用于确定科学数据与数据负责人的对应关系，保障科学数据的回溯和追踪。

5 总结

科学数据是科技创新和经济发展中不可或缺的基础性资源。在科学研究范式变革、大数据时代发展等因素的推动下，科学数据对科学研究的重要意义日益凸显。科学基金资助机构作为主要的科研项目资助和管理实体，需要提高科学数据管理水平，推动科学数据的开放共享。本文从数据生命周期的各阶段入手，提出联结和推动数据生命周期健康发展的科学数据管理的关键环节，包括：DMP 的制定与实施、科学数据汇交管理、科学数据开放共享和科学数据的可持续维护。其中，DMP 作为科学数据管理的纲领性文件，伴随科学数据生命周期的各个阶段。科学数据汇交管理应以发挥科学数据的最大价值为目标；应通过数据库的建立和数据汇交流程和模式的设计，为数据的特色化汇交和便捷共享打下良好的软、硬件基础。科学数据开放共享延续科学数据汇交的管理体系，通过多方共同参与的共享

机制提高用户和数据共享中心的良性互动。同时，为延长科学数据寿命，应对科学数据进行可持续维护，通过数据唯一标识、数据溯源、优化数据存储等手段，最大化科学数据价值，以持续推进科学数据开放共享，不断增强科技创新能力。

参考文献

1 温亮明, 张丽丽, 黎建辉. 大数据时代科学数据共享伦理问题研究. 情报资料工作, 2019, 40(2): 38-44.

2 尤霞光, 盛小平. 8个国际组织科学数据开放共享政策的比较与特征分析. 情报理论与实践, 2017, 40(12): 40-45.

3 Borgman C L. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 2012, 63(6): 1059-1078.

4 李正超. 国内科学数据共享平台建设现状及发展策略研究. 图书馆理论与实践, 2018, (8): 108-112.

5 Kim C W, Yoon H, Jin D, et al. Integrated management system for a large computing resources in a scientific data center. The Journal of Supercomputing, 2016, 72(9): 3511-3521.

6 张丽丽, 温亮明, 石蕾, 等. 国内外科学数据管理与开放共享的最新进展. 中国科学院院刊, 2018, 33(8): 774-782.

7 D'Anca A, Conte L, Nassisi P, et al. A multi-service data management platform for scientific oceanographic products. Natural Hazards and Earth System Sciences, 2017, 17(2): 171-184.

8 郁林羲. 全球开放获取运动中开放获取模型探析. 科技与出版, 2020, (8): 109-117.

9 Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 2016, 3: 160018.

10 Barisits M, Beermann T, Berghaus F, et al. Rucio: Scientific data management. Computing and Software for Big Science, 2019, 3(1): 1-19.

11 周满英, 付禄. 数据策管生命周期模型比较研究. 图书馆研究与工作, 2018, (9): 34-37.

12 Weatherburn J. Managing and sharing research data: A guide to good practice. The Australian Library Journal, 2016, 65(2): 135-136.

13 Li Y, Kennedy G, Ngoran F, et al. An ontology-centric architecture for extensible scientific data management systems. Future Generation Computer Systems, 2013, 29(2): 641-653.

14 Tedersoo L, Küngas R, Oras E, et al. Data sharing practices and data availability upon request differ across scientific disciplines. Scientific Data, 2021, 8: 192.

15 卫军朝, 张春芳. 国内外科学数据管理平台比较研究. 图书情报知识, 2017, (5): 97-107.

Scientific Data Management from Perspective of Scientific Funding Agencies

ZHAO Qiuhong* LI Yuanrui DENG Xiuquan ZHANG Chu ZHANG Baofeng

(School of Economics and Management, Beihang University, Beijing 100191, China)

Abstract With the development of technologies such as the acquisition, storage, analysis, and processing of scientific data, scientific research and innovation are gradually moving towards the era of big data, which takes scientific data as the basic scientific and technological resources. Moreover, the data-driven research paradigm has been widely used in the practical work of various disciplines, and the value of scientific data has increased to a prominent position of scientific research and innovation. As a result, the scientific data management responsibilities of research funding agencies are becoming increasingly important. Therefore, based on the analysis of the driving factors of the demand for scientific data management, this paper reviews the current experience of scientific data management practice in developed countries, and points out that scientific data management activities should connect all stages of the whole life cycle of scientific data, so as to extend the life cycle, to expand the value, and to promote its healthy and sustainable development. Therefore, this paper puts forward the management strategies for the whole life cycle of scientific data, including the formulation and implementation of scientific data management plan, scientific data collection management, open sharing of scientific data, and sustainable maintenance of scientific data. Then, implementation suggestions related to these strategies are put forward.

Keywords scientific data, scientific funding agencies, data collection, data sharing



赵秋红 北京航空航天大学学术委员会委员、经济管理学院教授、博士生导师。中国系统工程学会常务理事。主要研究领域：复杂系统管理与优化。E-mail: qhzhao@buaa.edu.cn

ZHAO Qiuhong Professor, Doctoral Supervisor of School of Economics and Management, Beihang University. She currently serves as Member of Academic Committee in Beihang University. She is also the executive member of the Systems Engineering Society of China. Her main research interest is the complex system management and optimization. E-mail: qhzhao@buaa.edu.cn

■责任编辑：张帆

*Corresponding author